Original Article

DOI: 10.4274/haseki.galenos.2025.79663
Med Bull Haseki 2025;63(3):123-127

Large Language Models and Male Circumcision: A Reliability Assessment

Ismail Ulus*, Gokhan Ceker**, Ibrahim Hacibey**

*University of Health Sciences Türkiye, Bagcilar Training and Research Hospital, Clinic of Urology, Istanbul, Türkiye
**University of Health Sciences Türkiye, Basaksehir Cam and Sakura City Hospital, Clinic of Urology, Istanbul, Türkiye

Abstract

Aim: Male circumcision remains routine in some countries for neonatal or religious reasons; however, it continues to be the subject of ongoing debate concerning its health benefits, potential risks, and implications for bodily autonomy. This study aims to evaluate the reliability of patient-facing content generated by four widely used large language models (LLMs) on various aspects of male circumcision.

Methods: A search regarding LLMs was conducted using 20 standardized questions on 10 May 2025. Responses from ChatGPT, Copilot, Gemini, and Perplexity were evaluated by three independent experts. Inter-rater reliability was assessed with the intraclass correlation coefficient, and model performance differences were analyzed using Kruskal-Wallis tests with Bonferroni correction.

Results: Inter-rater reliability was strong, with an intraclass correlation coefficient of 0.79 ($p<0.001$). Perplexity demonstrated statistically significant lower performance compared to ChatGPT, Copilot, and Gemini when evaluated across the thematic domains ($p<0.001$). Similarly, Perplexity performed statistically significantly worse than the other models across the criteria of clarity, structure, utility, and factual accuracy ($p<0.001$).

Conclusion: Gemini and Copilot were the top performers across both thematic domains and evaluation criteria, highlighting substantial differences among LLMs in their ability to provide accurate and well-structured medical information regarding male circumcision. While ChatGPT shows promise for patient guidance, the inconsistent performance of models such as Perplexity highlights the need for cautious implementation and continuous oversight in healthcare communication.

Keywords: Male circumcision, large language models, patient education

Introduction

Male circumcision, one of the oldest and most widely performed surgical procedures worldwide, holds significant cultural, religious, and medical importance (1). While deeply rooted in tradition for many communities, the medical indications and ethical implications of male circumcision remain subjects of ongoing debate and divergent perspectives within the healthcare community (2).

The digital age has ushered in an era in which patients increasingly seek health-related information through online resources, a trend further amplified by the rapid emergence of large language models (LLMs) that offer seemingly instant guidance (3). From chatbots delivering immediate responses to search engine algorithms curating vast repositories of medical content, artificial intelligence (AI) is rapidly emerging as a powerful intermediary in patient education and healthcare decision-making (4). However, this growing reliance on LLM-driven information raises critical concerns, particularly in the context of sensitive and frequently debated topics such as male circumcision, where misinformation can significantly impact individual health decisions and overall well-being (5).

This article aims to address this pressing concern by critically evaluating the reliability of AI powered patient guidance on male circumcision. We hypothesized that

Corresponding Author: Ismail Ulus, MD, University of Health Sciences Türkiye, Bagcilar Training and Research Hospital, Clinic of Urology, Istanbul, Türkiye
E-mail: ismailulus06@yahoo.com **ORCID:** orcid.org/0000-0002-2005-9734
Received: 25.06.2025 **Accepted:** 12.08.2025 **Publication Date:** 29.08.2025

Cite this article as: Ulus I, Ceker G, Hacibey I. Large language models and male circumcision: a reliability assessment. Med Bull Haseki. 2025;63(3):123-127

LLMs may be insufficient in providing accurate information on male circumcision and that significant performance differences may exist among LLMs. This study evaluates the accuracy, completeness, and potential biases embedded in the information provided by various AI platforms.

Materials and Methods

Compliance with Ethical Standards

As this study is based solely on AI-generated responses and does not involve human participants or the use of personal data, ethical committee approval was waived.

Study Design and Questionnaire

This cross-sectional study evaluated the responses generated by four AI models, ChatGPT (GPT-4.5), Microsoft Copilot, Google Gemini 2.5, and Perplexity AI, in response to a set of 20 questions on 10 May 2025. These questions, covering the medical, cultural, and psychological aspects of male circumcision, are primarily based on the technical report of the American Academy of Pediatrics Task Force on Circumcision (Table 1) (6). To minimize search bias associated with user history and personalized content, the searches were conducted without signing into AI platform accounts and using the incognito mode of the Google Chrome web browser. Responses were assessed and compared by three independent experts with extensive knowledge and experience in the field, using predefined evaluation criteria.

Table 1. List of questions

Medical indications and benefits
1. What are the medical benefits of male circumcision? 2. Is circumcision effective in preventing urinary tract infections? 3. Does circumcision reduce the risk of HIV or other sexually transmitted infections? 4. Can circumcision prevent penile cancer?
Risks and complications
5. What are the potential risks and complications of circumcision? 6. Is circumcision painful? How is pain managed during and after the procedure? 7. Can circumcision lead to erectile dysfunction or sexual problems? 8. What are the signs of complications after circumcision?
Myths and misinformation
9. Does circumcision cause infertility? 10. Can circumcision reduce sexual pleasure or sensation? 11. Is it true that circumcision makes the penis longer? 12. Can the foreskin grow back after circumcision?
Lifestyle and patient concerns
13. How long is the recovery after adult circumcision? 14. When can I resume sexual activity after circumcision? 15. What kind of daily hygiene is needed after circumcision? 16. Is adult circumcision safe and common?
Pediatric and cultural aspects
17. Should newborns be circumcised? What are the pros and cons? 18. What are the cultural or religious reasons for circumcision? 19. Is it ethical to circumcise children who cannot consent? 20. Are there any alternatives to circumcision for medical conditions like phimosis?

HIV: Human immunodeficiency virus

The questions were categorized into five thematic domains: "Medical Indications and Benefits", "Risks and Complications", "Myths and Misinformation", "Lifestyle and Patient Concerns", and "Pediatric and Cultural Aspects". Each thematic domain comprises four questions that comprehensively explore the topic of male circumcision from both medical and social perspectives.

Evaluation of Responses

The responses generated by the AI models were evaluated based on the following criteria:

Relevance: the extent to which the response was appropriate and aligned with established medical evidence,

Clarity: the degree to which the information was clearly and understandably communicated.

Structure: the logical organization and coherence of the response, Utility: the usefulness of the information for patient education and general guidance.

Factual Accuracy: the consistency of the response with verified medical facts. Each criterion was assessed using a five-point Likert scale, with scores ranging from 1 (lowest) to 5 (highest).

Statistical Analysis

Statistical analyses were conducted using SPSS software, version 29.0 (IBM Corp., Armonk, NY, USA). The normality of variable distributions was evaluated using the Kolmogorov-Smirnov and Shapiro-Wilk tests, complemented by visual assessments through quantile-quantile plots and histograms. Comparisons of scores among the AI groups were performed using the Kruskal-Wallis test. For variables showing statistically significant differences, post-hoc pairwise comparisons were conducted using the Bonferroni correction to adjust for multiple testing. A p-value of <0.05 was considered statistically significant for all analyses.

Results

Inter-rater reliability among the three independent expert reviewers was assessed using a two-way mixed-effects model based on absolute agreement. The analysis yielded an intraclass correlation coefficient of 0.79 (95% confidence interval: 0.711-0.868), indicating good inter-rater agreement among the reviewers ($p<0.001$).

When overall thematic scores were aggregated, Gemini and Copilot emerged as the top performers, each achieving a mean score of 4.65. ChatGPT performed moderately with a mean score of 4.40, while Perplexity consistently underperformed, recording a total mean score of 4.06. These differences were statistically significant ($p<0.001$)

Table 2. Comparative evaluation of AI model performance across thematic domains					
	ChatGPT	**Gemini**	**Copilot**	**Perplexity**	**p-value**
Medical indications and benefits	4.40±0.38	4.53±0.44	4.25±0.90	4.40±0.48	0.798
Risks and complications	4.40±0.45^{a,c}	4.48±0.44^{a}	4.80±0.25^{b}	3.95±0.54^{c}	<0.001
Myths and misinformation	4.43±0.57^{a,b}	4.68±0.37^{a}	4.78±0.26^{a}	4.10±0.66^{b}	0.003
Lifestyle and patient concern	4.40±0.45^{a}	4.65±0.29^{a}	4.63±0.22^{a}	3.98±0.60^{b}	<0.001
Pediatric and cultural aspects	4.38±0.39^{a}	4.90±0.21^{b}	4.80±0.25^{b}	3.88±0.65^{a}	<0.001
Total	4.40±0.44^{a}	4.65±0.38^{b}	4.65±0.49^{b}	4.06±0.60^{c}	<0.001

Different superscript letters indicate statistical significance between groups
AI: Artificial intelligence

(Table 2).

Within the medical indications and benefits domain, Gemini obtained the highest mean score, while Copilot recorded the lowest; however, the differences among the models were not statistically significant (p=0.798) (Figure 1). Copilot received the highest scores in the risks and complications domain, followed by Gemini and ChatGPT, while Perplexity demonstrated the weakest performance, receiving significantly lower ratings compared to the other models (p<0.001). In the myths and misinformation domain, Copilot and Gemini outperformed ChatGPT and Perplexity (p=0.003).

In the lifestyle and patient concerns domain, Gemini and Copilot demonstrated strong and comparable performance, whereas Perplexity exhibited a marked decline. The differences among the models in this domain were statistically significant (p<0.001). Performance in the pediatric and cultural aspects domain also varied significantly across models, with Gemini achieving the highest score, closely followed by Copilot, while ChatGPT and, particularly, Perplexity received lower evaluations in this area (p<0.001).

In addition to the thematic domains, the models were evaluated based on five global evaluation criteria: relevance, clarity, structure, utility, and factual accuracy (Figure 2). Under the clarity criterion, Gemini and Copilot were rated as the most comprehensible, while Perplexity received significantly lower scores, indicating weaker performance in effectively conveying information (p<0.001) (Table 3). In terms of structural coherence, Copilot achieved the highest ratings for organization and logical flow, followed closely by Gemini, while Perplexity was rated significantly lower, indicating marked differences across the models (p<0.001).

The utility of the responses varied significantly among the models (p<0.001), with Gemini and Copilot providing the most practically useful information for patient guidance, while ChatGPT and Perplexity received lower utility scores. Perplexity received the significantly lowest rating for factual accuracy, highlighting concerns about the reliability of its responses in conveying accurate medical information (p<0.001).

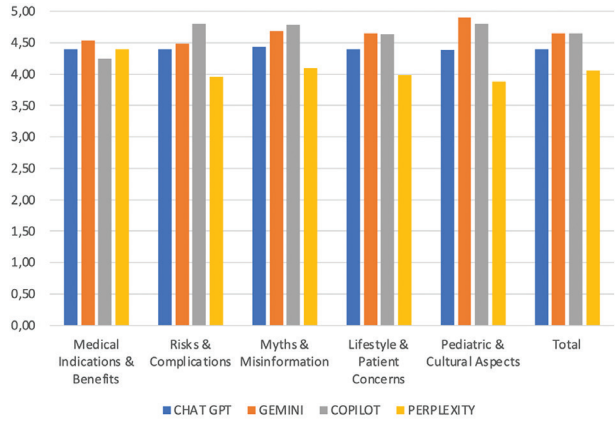

Figure 1. Mean scores of AI models across thematic domains
AI: Artificial intelligence

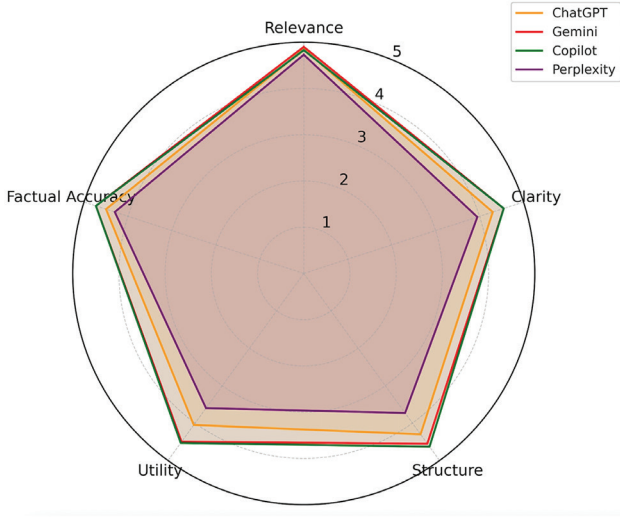

Figure 2. Radar chart comparing the performance of AI models across evaluation criteria
AI: Artificial intelligence

Table 3. Comparative evaluation of AI model performance across evaluation criteria					
	ChatGPT	**Gemini**	**Copilot**	**Perplexity**	**p-value**
Relevance	4.85±0.24	4.90±0.26	4.83±0.47	4.73±0.38	0.279
Clarity	4.30±0.34[a]	4.55±0.43[a]	4.55±0.43[a]	3.95±0.36[b]	<0.001
Structure	4.30±0.38[a]	4.55±0.36[a]	4.63±0.46[a]	3.73±0.53[b]	<0.001
Utility	4.05±0.46[a]	4.50±0.40[b]	4.53±0.53[b]	3.60±0.50[a]	<0.001
Factual accuracy	4.50±0.36[a]	4.73±0.34[a]	4.73±0.57[a]	4.3±0.47[b]	<0.001
Total	4.40±0.44[a]	4.65±0.38[b]	4.65±0.49[b]	4.06±0.60[c]	<0.001
Different superscript letters indicate statistical significance between groups AI: Artificial intelligence					

Discussion

This study presents a comprehensive comparative analysis of four LLMs in the generation of patient-facing educational content on male circumcision. The findings reveal that, although all models generated responses that were topically relevant, their performance varied substantially across the domains of clarity, structural organization, practical utility, and factual accuracy. Gemini and Copilot emerged as the most consistent and reliable in conveying medically and culturally sensitive information, whereas Perplexity demonstrated significantly lower performance across these domains.

The reliability of our expert evaluation was supported by a strong inter-rater agreement, consistent with established standards in LLM benchmarking and health communication research. While all models demonstrated high relevance scores, relevance alone proved to be an insufficient indicator of overall communication quality. Recent studies have shown that LLMs are prone to generating "hallucinated" content, plausible-sounding yet incorrect or fabricated information, particularly in complex and high-stakes domains such as medicine (7). This limitation underscores the importance of robust evaluation frameworks that extend beyond surface-level relevance. Effective patient education relies not only on topical alignment but also on the accuracy, structural clarity, and practical utility of the information and criteria that only a subset of the models in our analysis consistently fulfilled.

Emerging literature further corroborates the variable performance of AI in healthcare communication, highlighting both its potential and its limitations in generating accurate, patient-centered dialogue. Huang et al. (4) examined the diagnostic capabilities of chatbots, identifying significant limitations in their ability to manage clinical uncertainty and complex case scenarios. Similarly, Menz et al. (5) highlighted the risks associated with AI-generated health misinformation, advocating for rigorous oversight in clinical settings to mitigate potential harms.

Our study found that Copilot particularly excelled in dispelling misconceptions and explaining complications, as evidenced by its strong performance in the risks and complications as well as the myths and misinformation domains. These findings align with those of Anisuzzaman et al. (8), who demonstrated that domain-specific fine-tuning and interface design substantially impact the performance of LLMs in detecting health misinformation.

Gemini's superior performance in the pediatric and cultural aspects domain is noteworthy and suggests the benefits of enhanced contextual training tailored to these areas. This finding is consistent with observations by Kung et al. (9), who reported that newer-generation LLMs outperform earlier versions such as GPT-3.5 in United States Medical Licensing Examination style medical reasoning, particularly in tasks requiring nuanced communication.

In contrast, Perplexity's suboptimal performance, particularly in the areas of structural organization, clarity, and factual accuracy, raises important concerns regarding its readiness for deployment in healthcare-related applications. This supports the concerns raised by Thorp (10), who emphasized the unpredictability and opacity of LLMs in generating clinical advice.

Although ChatGPT demonstrated moderate performance, particularly in terms of factual accuracy, it lagged Copilot and Gemini in overall utility and the quality of patient-oriented communication. This discrepancy aligns with prior research indicating that even advanced LLMs often struggle to balance clinical precision with readability and empathetic tone in patient-facing communications (3). Empathetic communication has been shown to significantly enhance patient trust and engagement, but current LLMs remain limited in their ability to simulate empathetic dialogue in a manner that is both medically appropriate and contextually sensitive (11).

Considerable confusion exists among patients and individuals seeking health-related information regarding the reliability of available AI-based sources, and the spread of misinformation may lead to adverse health outcomes. Developing open-access health information content based on standardized guidelines, along with its classification according to specific purposes, is critically important. Future research may focus on designing alternative platforms to address this need.

Study Limitations

Several limitations warrant consideration. The evaluation was restricted to English-language content and did not incorporate assessments of emotional tone or potential biases in the models' outputs. Moreover, responses were evaluated in a controlled setting rather than through real-time user interactions, which may limit the ecological validity of our findings. Additionally, we did not examine the impact of AI-generated responses on patient decision-making, an important area that warrants future investigation. Given that algorithmic guidance can significantly influence user decision-making, even when the information provided is flawed or overly simplistic, further investigation into the behavioral impact of AI-generated content is warranted (12). The integration of intelligent systems into healthcare delivery will inevitably necessitate the development of new frameworks for accountability, transparency, and ethical oversight (13).

Conclusion

This investigation underscores the critical importance of evaluating the role of AI in shaping patient understanding of male circumcision in the digital era. Our assessment reveals that, although AI tools provide accessible medical information, their reliability remains variable. Gemini and Copilot demonstrated significantly superior performance across both thematic domains and evaluation criteria, whereas Perplexity lagged behind in all assessments. While AI can serve as a valuable adjunct for medical guidance, it should not replace the clinical judgment of healthcare professionals. Further research is needed to explore the broader implications of AI in healthcare and to develop strategies for its responsible and ethical deployment.

Ethics

Ethics Committee Approval: As this study is based solely on AI-generated responses and does not involve human participants or the use of personal data, ethical committee approval was waived.

Informed Consent: Since this study is based solely on responses generated by artificial intelligence and does not include human participant data or direct patient information, patient consent was not required.

Footnotes

Authorship Contributions

Concept: G.C., Design: G.C., I.H., Data Collection or Processing: I.U., I.H., Analysis or Interpretation: I.U., I.H., Literature Search: I.U., G.C., Writing: I.U., I.H.

Conflict of Interest: No conflicts of interest were declared by the authors.

Financial Disclosure: This study received no financial support.

References

1. Warees WM, Anand S, Leslie SW, Rodriguez AM. Circumcision. 2024 May 2. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025.
2. Wheeler R. Non-therapeutic circumcision of boys: a family matter. Arch Dis Child. 2025;11:417-8.
3. Mesko B, Győrffy Z. The rise of the empowered physician in the digital health era: viewpoint. J Med Internet Res. 2019;21:e12490.
4. Huang RS, Benour A, Kemppainen J, Leung FH. The future of AI clinicians: assessing the modern standard of chatbots and their approach to diagnostic uncertainty. BMC Med Educ. 2024;24:1133.
5. Menz BD, Modi ND, Sorich MJ, Hopkins AM. Health disinformation use case highlighting the urgent need for artificial intelligence vigilance: weapons of mass disinformation. JAMA Intern Med. 2024;184:92-6.
6. American Academy of Pediatrics Task Force on Circumcision. Male circumcision. Pediatrics. 2012;130:e756-85.
7. Yi Y, Kim KJ. The feasibility of using generative artificial intelligence for history taking in virtual patients. BMC Res Notes. 2025;18:80.
8. Anisuzzaman DM, Malins JG, Friedman PA, Attia ZI. Fine-tuning large language models for specialized use cases. Mayo Clin Proc Digit Health. 2024;3:100184.
9. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198.
10. Thorp HH. ChatGPT is fun, but not an author. Science. 2023;379:313.
11. Chen D, Chauhan K, Parsa R, Liu ZA, Liu FF, Mak E, et al. Patient perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer. NPJ Digit Med. 2025;8:275.
12. Ghanvatkar S, Rajan V. Evaluating explanations from AI algorithms for clinical decision-making: a social science-based approach. IEEE J Biomed Health Inform. 2024;28:4269-80.
13. Bidenko NV, Stuchynska NV, Palamarchuk YV, Matviienko MM. Integrating artificial intelligence in healthcare practice: challenges and future prospects. Wiad Lek. 2025;78:1199-205.